Reprinted from *J. Mol. Biol.* (1999) **000**, 000–000

JMB

The Relationship Between Protein Structure and Function: A Comprehensive Survey Focusing on Enzymes

Hedi Hegyi and Mark Gerstein

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1999		**00-00-1999 to 00-00-1999**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Relationship Between Protein Structure and Function: A Comprehensive Survey Focusing on Enzymes	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Yale University ,Department of Molecular & Biochemistry ,266 Whitney Avenue,New Haven ,CT,06520**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**19**	

Article No. jmbi.1999.2661 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 000, 1–18

JMB

The Relationship Between Protein Structure and Function: A Comprehensive Survey Focusing on Enzymes

Hedi Hegyi and Mark Gerstein*

Department of Molecular Biophysics & Biochemistry Yale University, 266 Whitney Avenue, PO Box 208114 New Haven, CT, 06520 USA

For most proteins in the genome databases, function is predicted *via* sequence comparison. In spite of the popularity of this approach, the extent to which it can be reliably applied is unknown. We address this issue by systematically investigating the relationship between protein function and structure. We focus initially on enzymes classified by the Enzyme Commission (EC) and relate these to structurally classified proteins in the SCOP database. We find that the major SCOP fold classes have different propensities to carry out certain broad categories of functions. For instance, alpha/beta folds are disproportionately associated with enzymes, especially transferases and hydrolases, and all-alpha and small folds with non-enzymes, while alpha + beta folds have an equal tendency either way. These observations for the database overall are largely true for specific genomes. We focus, in particular, on yeast, analyzing it with many classifications in addition to SCOP and EC (i.e. COGs, CATH, MIPS), and find clear tendencies for fold-function association, across a broad spectrum of functions. Analysis with the COGs scheme also suggests that the functions of the most ancient proteins are more evenly distributed among different structural classes than those of more modern ones. For the database overall, we identify the most versatile functions, i.e. those that are associated with the most folds, and the most versatile folds, associated with the most functions. The two most versatile enzymatic functions (hydro-lvases and *O*-glycosyl glucosidases) are associated with seven folds each. The five most versatile folds (TIM-barrel, Rossmann, ferredoxin, alpha-beta hydrolase, and P-loop NTP hydrolase) are all mixed alpha-beta structures. They stand out as generic scaffolds, accommodating from six to as many as 16 functions (for the exceptional TIM-barrel). At the conclusion of our analysis we are able to construct a graph giving the chance that a functional annotation can be reliably transferred at different degrees of sequence and structural similarity. Supplemental information is available from http://bioinfo.mbb.yale.edu/genome/foldfunc.

*Corresponding author

Keywords: Author supply keywords please

Introduction

The problem of determining function from sequence

An ultimate goal of genome analysis is to determine the biological function of all the gene products in a genome. However, the function of only a minor fraction of proteins has been studied experimentally, and, typically, prediction of function is based on sequence similarity with proteins of known function. That is, functional annotation is transferred based on similarity. Unfortunately, the relationship between sequence similarity and functional similarity is not as straightforward. This has been commented on in numerous reviews (Bork & Koonin, 1998; Karp, 1998). Karp (1998), in particular, has noted that transferring of incorrect functional information threatens to

Abbreviations used: EC, Enzyme Commission; ORF open reading frame.

E-mail address of the corresponding author: Mark.Gerstein@yale.edu

0022-2836/99/000001–18 $30.00/0

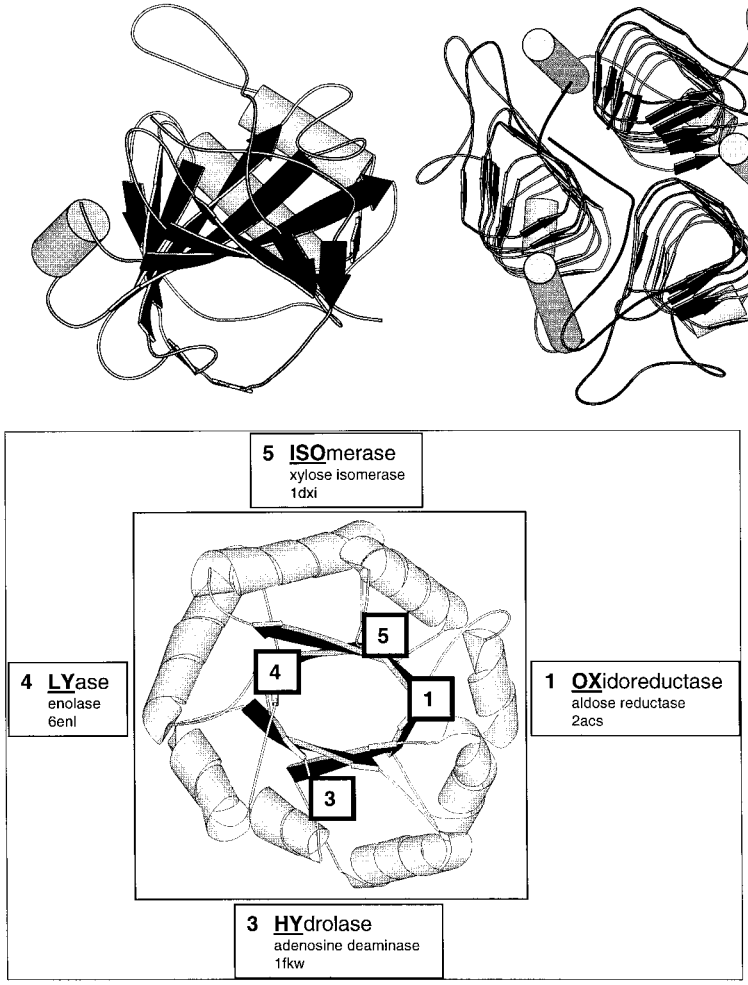


Figure 1. Specific example of convergent and divergent evolution. Top, an example of convergent evolution, showing structures of two carbonic anhydrases with the same enzymatic function (EC number 4.2.1.1), but with different folds. The Figure was drawn with Molscript (Kraulis, 1991) from 1THJ (left-handed beta helix) and 1DMX (flat beta sheet). Bottom, an example of possible divergent evolution, the TIM-barrel. This fold functions as a generic scaffold catalyzing 15 different enzymatic functions. A schematic Figure of the TIM-barrel fold is shown with numbers in boxes indicating the different location of the active site in four proteins that have this fold. These four proteins, xylose isomerase, aldose reductase, enolase, and adenosine deaminase, carry out very different enzymatic functions, in four of the main EC classes (1.*.*, 3.*.*, 4.*.*, and 5.*.*). They have active sites at very different locations in the barrel, yet they all share the same fold.

progressively corrupt genome databases through the problem of accumulating incorrect annotations and using them as a basis for further annotations, and so on.

It is known that sequence similarity does confer structural similarity. Moreover, there is a well-established quantified relationship between the extent of similarity in sequence and that in structure. First investigated by Chothia & Lesk (1986) the similarity between the structures of two proteins (in terms of RMS) appears to be a monotonic function of their sequence similarity. This fact is often exploited when two sequences are declared related, based on a database search by programs such as BLAST or FastA (Altschul *et al.*, 1997; Pearson, 1996). Often, the only common element in two distantly related protein sequences is their underlying structures, or folds.

Transitivity requires that the well-established relationship between sequence and structure, and the more indefinite one between sequence and function, imply an indefinite relationship between structure and function. Several recent papers have highlighted this, analyzing individual protein superfamilies with a single fold but diverse functions. Examples include the aldo-keto reductases, a large hydrolase superfamily, and the thiol protein esterases. The latter include the eye-lens and corneal crystallins, a remarkable example of functional divergence (Bork & Eisenberg, 1998; Bork *et al.*, 1994; Cooper *et al.*, 1993; Koonin & Tatusov, 1994; Seery *et al.*, 1998).

There are also many classic examples of the converse: the same function achieved by proteins with completely different folds. For instance, even though mammalian chymotrypsin and bacterial subtilisin have different folds, they both function as serine proteases and have the same Ser-Asp-His catalytic triad. Other examples include sugar kinases, anti-freeze glycoproteins, and lysyl-tRNA synthetases (Bork *et al.*, 1993; Chen *et al.*, 1997; Doolittle, 1994; Ibba *et al.*, 1997a,b).

Figure 1 shows well-known examples of each of these two basic situations: the same fold but different function (divergent evolution) and the same function but different fold (convergent evolution).

Protein classification systems

The rapid growth in the number of protein sequences and three-dimensional structures has made it practical and advantageous to classify proteins into families and more elaborate hierarchical systems. Proteins are grouped together on the

basis of structural similarities in the FSSP (Holm & Sander, 1998), CATH (Orengo *et al*., 1997), and SCOP databases (Murzin *et al*., 1995). SCOP is based on the judgments of a human expert FSSP, on automatic methods, and CATH, on a mixture of both. Other databases collect proteins on the basis of sequence similarities to one another, e.g. PROSITE, SBASE, Pfam, BLOCKS, PRINTS and ProDom (Attwood *et al*., 1998; Bairoch *et al*., 1997; Corpet *et al*., 1998; Fabian *et al*., 1997; Henikoff *et al*., 1998; Sonnhammer *et al*., 1997). Several collections contain information about proteins from a functional point of view. Some of these focus on particular organisms, e.g. the MIPS functional catalogue and YPD for yeast (Mewes *et al*., 1997; Hodges *et al*., 1998) and EcoCyc and GenProtEC for *Escherichia coli* (Karp *et al*., 1998; Riley, 1997). Others focus on particular functional aspects in multiple organisms, e.g. the WIT and KEGG databases which focus on metabolism and pathways (Selkov *et al*., 1997; Ogata *et al*., 1999), the ENZYME database which focuses obviously enough on enzymes (Bairoch, 1996), and the COGs system which focuses on proteins conserved over phylogenetically distinct species (Tatusov *et al*., 1997). The ENZYME database, in particular, contains all the enzyme reactions that have an Enzyme Commision (EC) number'' assigned in accordance with the International Nomenclature Committee and is cross-referenced with Swissprot (Bairoch, 1996; Bairoch & Apweiler, 1998; Barrett, 1997).

Our approach: systematic comparison of proteins classified by structure with those classified by function

One of the most valuable operations one can do to these individual classification systems is to cross-reference and cross-tabulate them, seeing how they overlap. We performed such an analysis here by systematically interrelating the SCOP, Swissprot and ENZYME databases (Bairoch, 1996; Bairoch & Apweiler, 1998; Murzin *et al*., 1995). For yeast we also have used the MIPS yeast functional catalogue, CATH, and COGs in our analysis. This enables us to investigate the relationship between protein function and structure in a comprehensive statistical fashion. In particular, we investigated the functional aspects of both divergent and convergent evolution, exploring cases where a structure gains a dramatically different biochemical function and finding instances of similar enzymatic functions performed by unrelated structures.

We concentrated on single-domain Swissprot proteins with significant sequence similarity to one of the SCOP structural domains. Since most of these proteins have a single assigned function, comparing them to individual structural domains, which can have only one assigned fold, allowed us to establish a one-to-one relationship between structure and function.

Recent related work

This work is following up on several recent reports on the relationship between protein structure and function. In particular, Martin *et al*. (1998) studied the relationship between enzyme function and the CATH fold classification. They concluded that functional class (expressed by top-level EC numbers) is not related to fold, since a few specific residues, not the whole fold, determine enzyme function. Russell (1998) also focused on specific side-chain patterns, arguing that these could be used to predict protein function. In a similar fashion, Russell *et al*. (1998) identified structurally similar ''supersites'' in superfolds. They estimated that the proportion of homologues with different binding sites, and therefore with different functions, is around 10%. In a novel approach, using machine learning techniques, des Jardins *et al*. (1997) predict purely from the sequence whether a given protein is an enzyme and also the enzyme class to which it belongs.

Our work is also motivated by recent work looking at whether or not organisms are characterized by unique protein folds (Frishman & Mewes, 1997; Gerstein, 1997, 1998a,b; Gerstein & Hegyi, 1998; Gerstein & Levitt, 1997). If function is closely associated with fold (in a one-to-one sense), one would think that when a new function arose in evolution, nature would have to invent a new fold. Conversely, if fold and function are only weakly coupled, one would expect to see a more uniform distribution of folds amongst organisms and a high incidence of convergent evolution. In fact, a recent study on microbial genome analysis claims that functional convergence is quite common (Koonin & Galperin, 1997). Another related paper systematically searched Swissprot for all such cases of what is termed ''analogous'' enzymes (Galperin *et al*., 1998).

Our work is also motivated by the recent work on protein design and engineering which aims to rationally change a protein function, for instance, to engineer a reporter function into a binding protein (Hellinga, 1997, 1998; Marvin *et al*., 1997).

Results

Overview of the 8937 single-domain matches

Our basic results were based on simple sequence comparisons between Swissprot and SCOP, the SCOP domain sequences being used as queries against Swissprot. We focused on ''mono-functional'' single-domain matches in Swissprot, i.e. those singe-domain proteins with only one annotated function. The detailed criteria used in the database searches are summarized in Materials and Methods.

Overall, a little more than a quarter of the proteins in Swissprot are enzymes, a similar fraction are of known structure, and about one-eighth are both. (More precisely, of the 69,113 analyzed pro-

teins in Swissprot, 19,995 are enzymes, 18,317 are structural homologues, and 8205 are both.) About half of the fraction of Swissprot that matched known structures were ''single-domain'' and about one-third of these were enzymes (8937 and 3359, respectively, of 18,317). We focus on these 8937 single-domain matches here. Notice how these numbers also show how the known structures are significantly biased towards enzymes: 45% (8205 out of 18,317) of all the structural homologues are enzymes *versus* 29% (19,995 out of 69,113) for all of Swissprot.

331 observed fold-function combinations

Figure 2 gives an overview of how the matches are distributed amongst specific functions and folds. The single-domain matches include 229 of the 361 folds in SCOP 1.35, and 91 of the 207 three-component enzyme categories in the ENZYME database (Bairoch, 1996). Each match combines a SCOP fold number on the structural side (columns in Figure 2) and a three-component EC category on the functional side (rows), with all the non-enzymatic functions grouped together into a single category with the artificial ''EC number'' of 0.0.0 (shown in the first row in Figure 2). This results in a table where each cell represents a potential fold-function combination. The table contains a maximum of 21,068 ($=229 \times 92$) possible fold-function combinations (and a minimum of 229 combinations, assuming only one function for every fold). We actually observe 331 of these combinations (1.6%, shown by the filled-in cells).

Overall, more than half of the functions are associated with at least two different folds, while less than half of the folds with enzymatic activity have at least two functions (51 out of 91 and 53 out of 128, respectively).

Summarizing the fold-function combinations by 42 broad structure-function classes

As listed in Table 1, folds can be subdivided in six broad fold classes (e.g. all-alpha, all-beta, alpha/beta, etc.). Likewise, functions can be broken into secen main classes, non-enzymes plus six enzyme classes, e.g. oxidoreductase, transferase, etc. This gives rise to 42 (6×7) structure-function classes. The way the 21,068 potential fold-function combinations are apportioned amongst the 42 classes is shown in Table 2A.

Table 2B shows the way the 331 observed combinations were actually distributed amongst the 42 classes. Comparing the number of possible combinations with that observed shows that the most densely populated region of the chart is the transferase, hydrolase and lyase functions in combi-

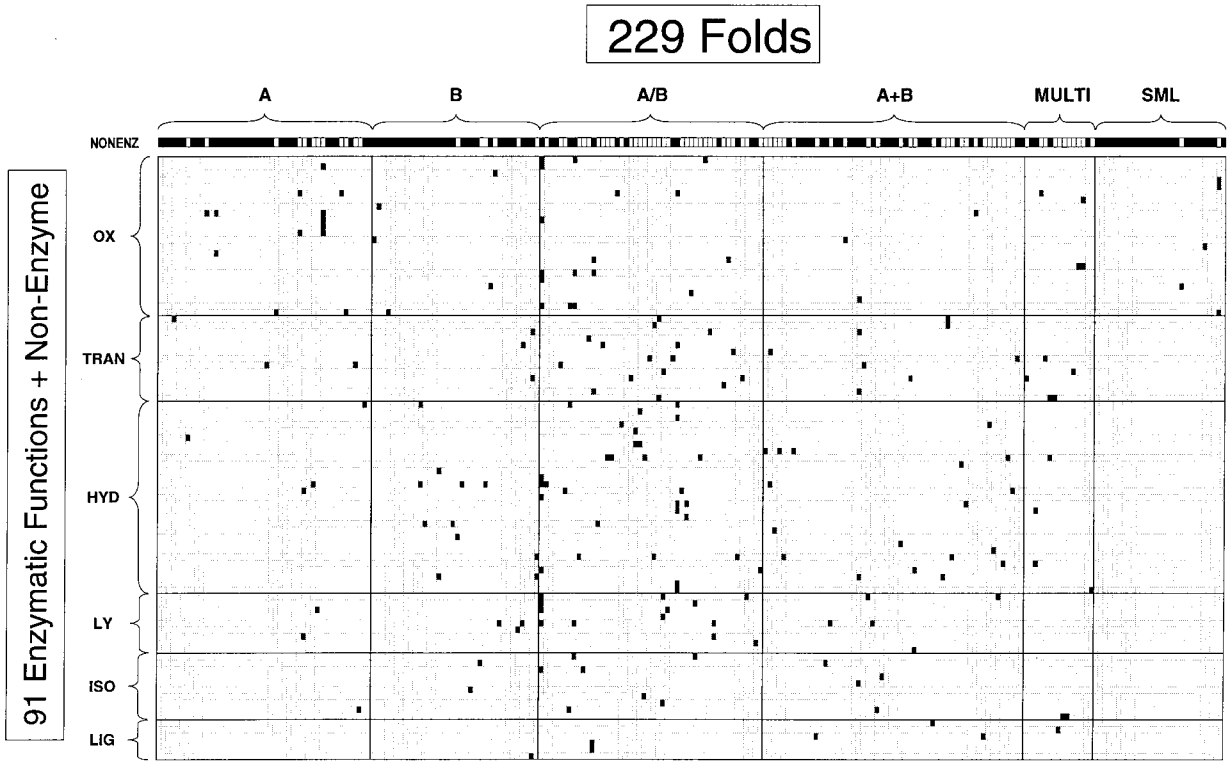


Figure 2. Overview of all the single-domain matches between proteins in Swissprot 35 and domains in SCOP 1.35. Sequences were compared with BLAST using the match criteria described in the methods. The matches are clustered into 92 functions (based on three-component EC numbers), which are arranged on each row, and 229 folds (based on SCOP fold numbers), which are arranged on each column. The first row indicates the matches with non-enzymes. There are, thus, 21,068 ($=92 \times 229$) possible combinations shown in the figure. Only the 331 are actually observed. These are indicated by filled squares.

Table 1. Broad structural and functional categories

A. *Functional categories in Swissprot 35*[a]

EC Category	Category name	Abbreviation	Num. of functions in category
0.0.0	Non-enzymes	NONENZ	1
1.*.*	Oxidoreductases	OX	86
2.*.*	Transferases	TRAN	28
3.*.*	Hydrolases	HYD	53
4.*.*	Lyases	LY	15
5.*.*	Isomerases	ISO	16
6.*.*	Ligases	LIG	9
		Total:	208

B. *Structural classes in SCOP 1.35*[b]

Fold class	Class name	Abbreviation	Num. of folds in class
1	All-alpha	A	81
2	All-beta	B	57
3	Alpha and beta	A/B	70
4	Alpha plus beta	A + B	91
5	Multi-domain	MULTI	19
6	Transmembrane	TM	9
7	Small proteins	SML	43
		Total:	361

[a] List of the functional (enzymatic) categories in Swissprot and the abbreviations used here. The values denote the number of three-component EC numbers in each category.

[b] List of the structural classes in SCOP studied here, and the abbreviations used for the classes. Values denote the number of folds in each class in SCOP 1.35. Class 6 is not used in the analysis.

nation with the alpha/beta fold class. This notion is in accordance with the general view that the most popular structures among enzymes fall into the alpha/beta class. In contrast, matches between small folds and enzymes are almost completely missing, except for five folds in the oxidoreductase category. There are also no all-alpha ligases and only one all-alpha isomerase.

Table 2C and D break down the 331 fold-function combinations in Table 2A into either just a number of folds or just a number of functions. That is, Table 2C lists the number of different folds associated with each of the 42 structure-function classes (corresponding to the non-zero columns in the relevant class in Figure 2), and Table 2D does the same thing for functions (non-zero rows in Figure 2). Comparing these tables back to the total number of combinations (Table 2A) reveals some interesting findings, keeping in mind that more functions than folds reveals probable divergence and that more folds than functions reveals probable convergence. For instance, the alpha/beta and alpha + beta fold classes contain similar numbers of folds, but the alpha/beta class has relatively more functions, perhaps reflecting a greater divergence. (Specifically, the alpha/beta class has 73 folds and 56 functions, while the alpha + beta class has 67 folds but only 35 functions.)

Table 2E shows the number of matching Swissprot sequences (from the total of 69,113) for each of the 42 structure-function classes. The most highly populated categories are the all-alpha non-enzymes, where 683 of the 1940 matches come from globins, and the all-beta non-enzymes, where 361 of the 1159 Swissprot sequences have matches with the immunoglobulin fold. These numbers are, obviously, affected by the biases in Swissprot. On the other hand, if we compare the total matches in Table 2E with the total combinations in Table 2B it is clear that the numbers do not directly correlate. For instance, fewer hydrolases in Swissprot have matches with alpha/beta folds than with alpha + beta folds (295 *versus*. 452), but the number of different combinations in the first case is 30, as opposed to only 18 in the second case. This suggests that our approach of counting combinations may not be as affected by the biases in the databanks as simply counting matches.

Table 2F and 2G give some rough indication of the statistical significance of the differences in the observed distribution of combinations. In Table 2F, using chi-squared statistics, we calculate for each individual structure class the chance that we could get the observed distribution of fold-function combinations over various functional classes if fold was not related to function. Then in table 2G, we reverse the role of fold and function, and calculate the statistics for each functional class.

Enzyme *versus* non-enzyme folds

On the coarsest level, function can be divided amongst enzymes and non-enzymes. Of the 229 folds present in Figure 2, 93 are associated only with enzymes and 101 are associated only with non-enzymes. The remaining folds were associated with both enzymatic and non-enzymatic activity. Finally, of the 93 purely enzymatic folds, 18 have multiple enzymatic functions.

Figure 3(a) shows a graphical view of the distribution of the different fold classes among these

broadest functional categories. The distribution is far from uniform. The all-alpha fold class has 30 non-enzymatic representatives, but only 12 purely enzymatic folds and four folds with ''mixed'' (both types of) functions. This implies that a protein with an all-alpha fold has *a priori* roughly twice the chance of having a non-enzymatic function over an enzymatic one. The all-beta fold class has six enzymatic, 17 non-enzymatic and 13 mixed folds. In the alpha/beta class, 34 folds are associated only with enzymes and five folds only with non-enzymes, whereas in the alpha + beta class this ratio is more balanced, 28 ''purely'' enzymatic folds *versus* 22 purely non-enzymatic ones.

Table 2. Statistics over 42 structure-function classes

A. *Number of possible combinations between folds and functions in each of 42 classes (number of cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	46	36	48	56	15	28	229
OX	1104	864	1152	1344	360	672	5496
TRAN	598	468	624	728	195	364	2977
HYD	1334	1044	1392	1624	435	812	6641
LY	414	324	432	504	135	252	2061
ISO	460	360	480	560	150	280	2290
LIG	276	216	288	336	90	168	1374
sum	4232	3312	4416	5152	1380	2576	21068

B. *Number of observed combinations between folds and functions in each of 42 classes (number of filled cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	13	5	17	3	4	5	47
TRAN	3	3	16	8	5		35
HYD	4	11	30	18	4		67
LY	2	3	13	5			23
ISO	1	2	7	4	2		16
LIG		1	2	3	1		7
sum	57	55	99	69	20	31	331

C. *Number of folds in each of the 42 classes (columns with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	7	5	9	3	3	3	30
TRAN	3	2	15	6	5		31
HYD	4	8	19	18	3		52
LY	2	3	8	5			18
ISO	1	2	7	4	2		16
LIG		1	1	3	1		6
sum	51	51	73	67	18	29	289

D. *Number of functions in each of the 42 classes (rows with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1	1	1	1	1	1	6
OX	8	5	9	3	3	5	33
TRAN	2	3	13	8	4		30
HYD	4	7	19	14	4		48
LY	2	2	7	3			14
ISO	1	2	5	4	1		13
LIG		1	2	2	1		6
sum	18	21	56	35	14	6	150

E. *Total number of matching Swissprot sequences in each of the 42 fold-function classes*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1940	1159	560	638	106	892	5295
OX	150	202	388	50	68	18	876
TRAN	65	14	363	116	174		732
HYD	116	394	295	452	92		1349
LY	40	47	168	104			359
ISO	2	54	122	22	2		202
LIG		5	26	69	24		124
sum	2313	1875	1922	1451	466	910	8937

F. *How much does each of the fold classes deviate from the average distribution of functions?*

	χ^2	*P*
A	17.5	<0.01
B	5.2	<0.6
A/B	32.5	<0.00002
A + B	7.7	<0.3
MULTI	9.9	<0.2
SML	27.8	<0.0002

continued

Table 2—*Continued*

G. *How much do each of the function classes deviate from the average distribution of folds?*

	χ^2	*P*
NONENZ	40.7	<0.0000002
OX	9.9	<0.08
TRAN	13.1	<0.03
HYD	17.3	<0.005
LY	10.2	<0.08
ISO	5.0	<0.5
LIG	4.3	<0.6

This Table shows various totals from Figure 2 distributed among the 42 structure-function classes i.e. the seven functional categories in Table 1A multiplied by the six structural categories in Table 1B. Part A shows how many potential fold-function combinations there are in Figure 2 amongst each of the 42 classes. Part B shows how many of these 21068 possible combinations are actually observed. Part C shows the total number of different folds (i.e. selected columns in Figure 1) in each class. Part D shows the total number of different functions (i.e. selected rows in Figure 2) in each class. Part E shows the total number of matching Swissprot proteins in the 42 classes. Note that to observe a fold-function combination one only needs the existence of a single match between a Swissprot protein and a SCOP domain. However, there can be many more. That is why the totals in this table sum up to so much larger an amount than 331.

Here is an example of how to read parts A to E of the table, focussing on the all-alpha, oxidoreductase region. Part A shows that there are 1104 cells, filled or unfilled, in this region, corresponding to possible combinations. Part B shows that 13 of these 1104 cells are filled, corresponding to observed all-alpha, oxidoreductase combinations. Part C shows that there are seven folds, corresponding to columns with filled cells in this region. Part D shows that there are eight functions, corresponding to rows with filled cells in this region. Finally, in part E we find that there are 150 Swissprot entries that have matches with a SCOP domain. They correspond to the 13 observed combinations in Part B.

Parts F and G give information on the statistical significance of the differences observed between the 42 structure-function classes. Part F gives the significance that the observed distribution of fold-function combinations in a given functional class is different than average (i.e. the null hypothesis that distribution of fold-function combinations is the same in each functional class). This is very similar to the derivation by Martin *et al*. (1998). A chi-squared statistic is computed for each of the seven functional classes in the conventional way: $\chi^2(f) = \Sigma_s\ (O_{sf} - E_{sf})^2/E_{sf}$, where for a given functional class f and structure class s, O_{sf} is the observed number of fold-function combinations and E_{sf} is the expected number. E_{sf} is simply computed from scaling the ''sum'' column and row in Part B of the table:$E_{sf} = T_sT_f/T$, where T_s is the total number of combinations in a given structural class s (sum row), T_f is the total number of combinations in a given functional class f (sum column), and T is the total observed number of combinations, 331. Part G gives the statistical significance that the observed distribution of fold-function combinations in a given structural class is different than average. To compute this one simply sums over functions instead of structures: $\chi^2(s) = \Sigma_f(O_{sf} - E_{sf})^2/E_{sf.}$ After each chi-squared statistic is reported, a rough probability or *P*-value is given. This gives the chance the observed distribution could be obtained randomly.

Restricting the comparison to individual genomes

Figure 3(a) applies to all of Swissprot. Figure 3(b) and (c) show the functional distribution of folds taking into account the matches only in two specific genomes, yeast and *E. coli*. Only a fraction of each genome could be taken into consideration for various reasons (156 proteins in yeast, 244 proteins in *E. coli*), mostly due to the great number of enzymes having multiple domains in both yeast and *E. coli*. Chi-squared tests show that the fold distribution in yeast does not differ significantly from that in Swissprot and that the one in *E. coli* differs only slightly ($P < 0.25$ and $P < 0.02$, respectively). The main difference between Swissprot and *E. coli* is the larger fraction of alpha/beta enzymatic folds in the latter (34/93 *versus* 26/49). There are also somewhat more non-enzymatic all-alpha and small folds in Swissprot than in the two genomes. This is principally due to the greater prevalence of globins, myosins, cytochromes, toxins, and hormones in Swissprot than in yeast and *E. coli*. Many of these, of course, are proteins usually associated with multicellular organisms. We did a preliminary version of the fold distribution for the worm *Caencrhaditis elegans*. As expected this distribution turns out to be similar to that of Swissprot (data not shown).

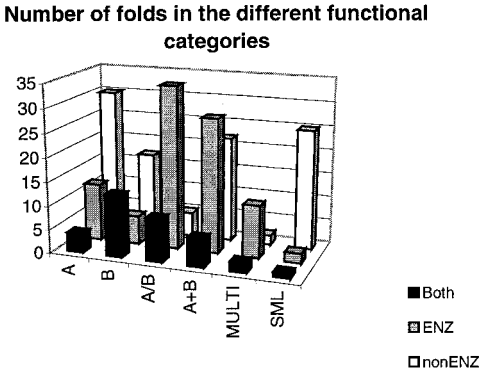
The yeast genome viewed from different classification schemes

In Figure 4 we focus on the yeast genome in more detail, trying to see the effect that different classification schemes have on our results. Although the total number of counts for our statistics decrease, in just using yeast relative to all of Swissprot, yeast provides a good reference frame to compare a number of classification schemes in as unbiased a fashion as possible. Also, yeast is one of the most comprehensively characterized organisms, and there are a number of functional classifications available exclusively for this organism.

In part Figure 4(a) we cross-tabulate the structure-function combinations in yeast using the SCOP and EC systems as we have done for all of Swissprot in Table 2B. The yeast distribution is fairly similar to that of Swissprot with the only major difference being somewhat more alpha/beta transferases and fewer alpha/beta hydrolases than expected. (A chi-squared test gives $P < \sim 0.05$ for the two distributions to differ. If either the transferase or hydrolase difference is removed, *P* increases to ~20 %.)

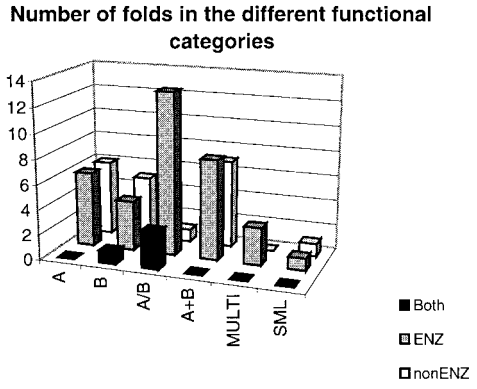
Figure 4(b) shows the structure-function combinations based on using the CATH structural classification (Orengo *et al*., 1997) instead of SCOP. For

A. All of Swissprot



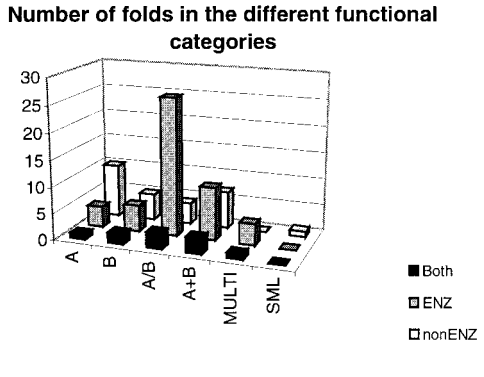
	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	4	13	9	6	2	1	35
ENZ	12	6	34	28	11	2	93
nonENZ	30	17	5	22	2	25	101

B. Yeast



	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	0	1	3	0	0	0	4
ENZ	6	4	13	8	3	1	35
nonENZ	6	5	1	7	0	1	20

C. E. coli



	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	1	2	3	3	1	0	10
ENZ	4	5	26	10	4	0	49
nonENZ	10	5	4	7	0	1	27

this Figure we mapped the SCOP classification of a yeast PDB match to its corresponding CATH classification and then cross-tabulated the structure-function combinations in the various classes. Essentially, this Figure shows the results reported by Martin *et al*. (1998) just for yeast.

In Figure 4(c) and (d), which show COGs *versus* SCOP cross-tabulations, we achieve the opposite of (b). We change the functional classifications scheme but keep SCOP for classifying structures. As was the case with the enzyme classification, but perhaps even more so, using COGs to classify function shows clearly that certain fold classes are associated with certain functions and *vice versa*. Most notably, whereas the functions associated with metabolism, which are mostly enzymes, are preferentially associated with the alpha/beta fold class, those associated with cellular processes (e.g. secretion) and information processing (e.g. transcription), show no such preference. They, in fact, show a marked preference for all-alpha structure. Small proteins are absent from most of the COGs classes, except one part of information processing and two in cellular processes.

The COGs system classifies functions for those proteins that have clear orthologues in different species. Thus, conclusions based on using yeast COGs should be readily applicable to other genomes. This point is highlighted in Figure 4(d), which shows a COGs *versus* SCOP classification for only the 110 COGs that are conserved across all the analyzed genomes (eight) and all three kingdoms. Thus, this sub-figure would appear *exactly* the same for *E. coli*, *Methanococcos jannaschii* or a number of other genomes. It clearly shows how much more common the information processing proteins are among the most conserved and ancient proteins. Moreover, note how these most ancient proteins appear to have less of a preference for a particular structural class than the ''more modern'' metabolic ones. This suggests that large-scale duplication of alpha/beta folds for use in metabolism is what gave rise to stronger fold-function association in Figure 3(c).

Figure 3. Chart with breakdown among structure-function classes in two genomes. Charts and Tables showing the number of folds in each fold class associated with only enzymatic (ENZ), only non-enzymatic (nonENZ), and both enzymatic and non-enzymatic functions (Both). The results are shown for (a) all of Swissprot, (b) for just the yeast genome, and (c) for just the *E. coli* genome. The results for individual domains in a minimum set of SCOP domains also support these tendencies (data not shown). The numbers in (b) are not based on the PSI-blast protocol used for Figure 4. Rather they are found just as ''subsets'' of the overall Swissprot results to make them readily comparable with the rest of the paper. Because of this the numbers in this figure will not match exactly those in Figure 4, the difference having to do with the greater number of fold-function combinations found by PSI-Blast as compared to WU-blast.

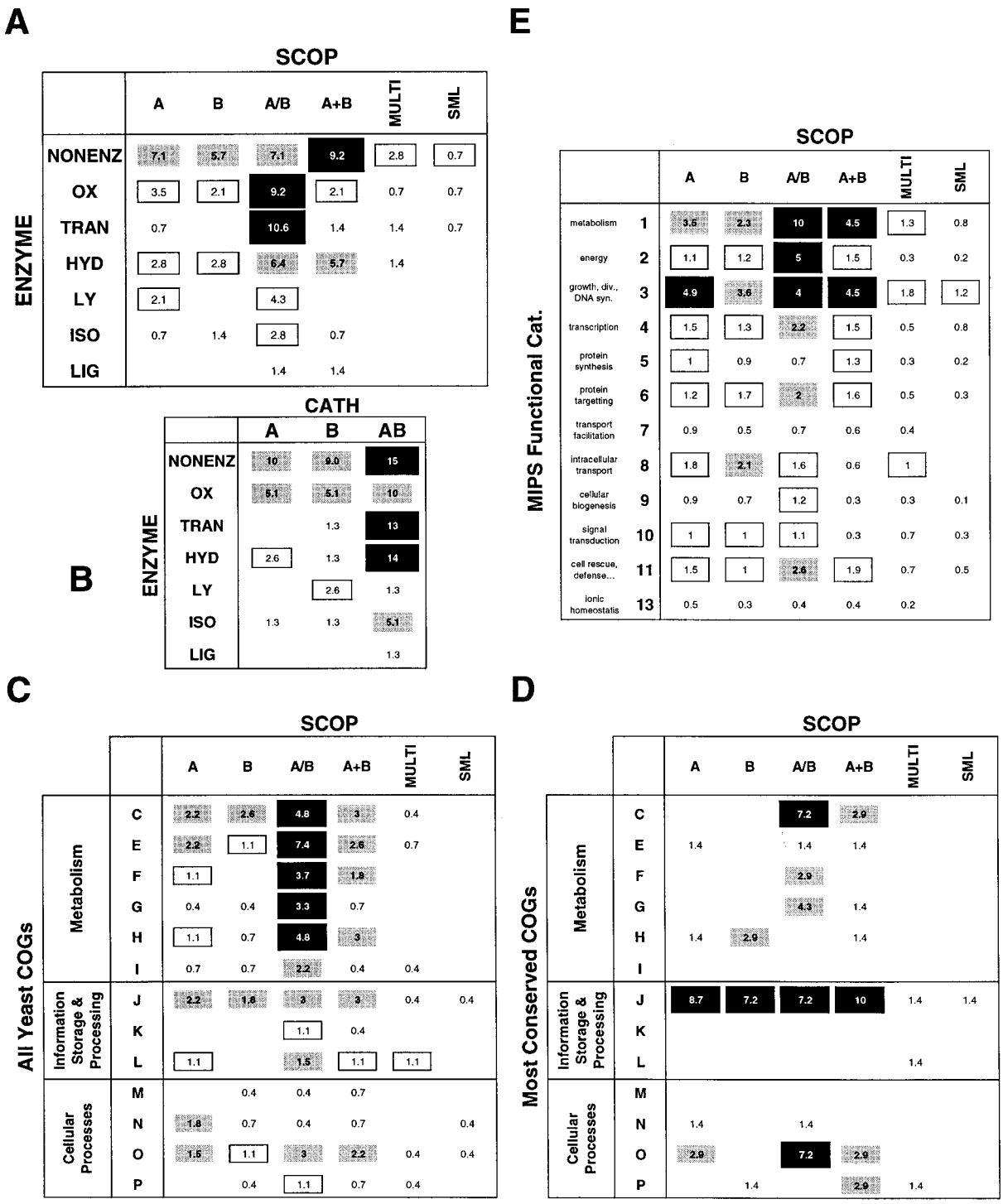


Figure 4. Structure-function classes in the yeast genome analyzed through a variety of classification schemes. This figure shows the distribution of fold function combinations in the yeast genome as analyzed by a variety of different structure and functional classifications. Each of the Figures is a cross-tabulation of one structural classification scheme (on the column heads) versus a functional classification (row heads). (a) SCOP *versus* ENZYME; (b) CATH *versus* ENZYME; (c) SCOP *versus* COGs; (d) SCOP *versus* Most Conversed COGs; (e) SCOP *versus* MIPS Functional Catalogue. Each of the grid boxes gives the number of fold-function combinations within a structure-function class. This number is expressed as a percentage of the total number of combinations in the diagram to make the graphs readily comparable. The total number of combinations in each of the sub figures is (a) 141, (b) 77, (c) 1207, (d) 120, and (e) 66. (a) and (e) is directly comparable with the cross tabulation in table 2B for all of Swissprot. In Parts D and E, we employ the COGs scheme in exactly the same fashion as we did the ENZYME classification. We form combinations between individual yeast COGs and SCOP folds (e.g. COG 0186 with fold 2.26) and then we place these combinations into larger structure-function classes. The COGs overall functional classes are denoted by a single letter and then are in turn grouped into three broader areas (so, for instance, the 0186-2.26 pair would go into the structure-function class all-beta, J). We, likewise, proceed similarly for the MIPS yeast functional catalogue. This gives each function a two or three component number similar to an EC number (e.g. 07.20.3 or 06.2). We use the first two numbers to create combinations with SCOP folds and then use the top number to create the functional classes shown in the diagram. For (e) we just use the 110 COGs that are present in all eight genomes in the current COGs analysis (*E. coli*, *H. influenzae*, *H. pylori*, *M. genitalium*, *M. pneumoniae*, *Synechocystis*, *M. jannaschii*, and yeast).

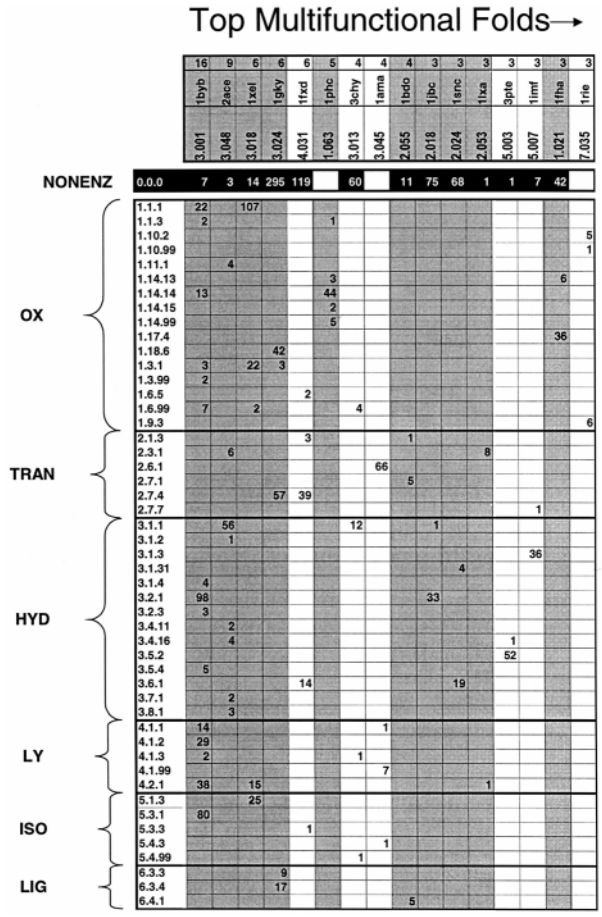


Figure 5. The most versatile folds. The functions associated with the 16 most versatile folds are shown. Values in the Table denote the number of matches between a particular fold type in pdb95d (designated by its fold number in SCOP 1.35) and an enzyme category (represented by the first three components of the respective EC numbers). Here and in the following Tables the same parameters were used for matching as in Figure 2. The numbers in the top row indicate the number of functions a particular fold is associated with. The identifiers above the fold numbers are either PDB or SCOP identifiers of representative structures (the latter only if the PDB entry contains more than one domain or chain). (See the legend to Table 3 for the syntax of SCOP identifiers.) The first row in the Table with the artificial 0.0.0 EC number shows the number of matches with non-enzymatic functions. Among the two all-alpha folds in the table, cytochrome P450 (1.063) is exclusively enzymatic, associated with five different enzyme functions, all related to cytochrome P450. Only one alpha + beta fold, ferredoxin (4.031), is present in the Table, predominantly with matches with non-enzymatic ferredoxins, but also with enzymes in four different enzyme classes. In the multi-domain class, beta-lactamase/D-ala carboxypeptidase (5.003) has the most matches with penicillinase (EC number 3.5.2) and only one match with a non-enzyme, which also binds penicillin but has no enzymatic activity (Coque *et al.*, 1993). The class of small domains is represented only with one fold, membrane-bound rubredoxin-like (7.035), and has matches only with enzymes. It is possible that some proteins classified as ''non-enzymes'' may indeed be enzymes, missing the corresponding EC number. In this case, our analysis may be potentially useful in pointing to which non-enzymes may actually be enzymes.

Figure 4(e) shows another functional classification scheme, the MIPS Yeast functional catalogue (Mewes *et al.*, 1997). Unlike the COGs scheme, this has the advantage of being applicable to every yeast open reading frame (ORF). However, it has many more categories and about a third of the yeast ORFs are classified into multiple categories (sometimes five or more), making interpretation of the results a bit more ambiguous.

The most versatile folds and the most versatile functions

Returning to considerations of all of Swissprot, Figure 5 lists the 16 most versatile folds. The top five are the TIM-barrel, the alpha-beta hydrolase fold, the Rossmann fold, the P-loop containing NTP hydrolase fold, and the ferredoxin fold. Four of these are alpha/beta folds and one is alpha + beta. All five have non-enzymatic functions as well as five to 15 enzymatic ones. The most versatile folds include four all-beta and two all-alpha folds.

Figure 6 lists the 18 functions that have the most different folds associated with them, each having at least three associated folds. The most versatile functions are those of glycosidases and carboxylyases (3.2.1 and 4.2.1), which are associated with seven different fold types each, recruited from at least three different fold classes. The next two most versatile functions, the phosphoric monoester hydrolases and the linear monoester hydrolases (3.1.3 and 3.5.1), are associated with six different fold types each. Most of the versatile functions are associated with folds in completely different fold classes. This suggests that these enzymes developed independently, providing many examples of convergent evolution. In contrast, only three functions, all oxidoreductases, are associated with folds in a single class (last three rows in Figure 6). These folds are all alpha/beta, namely the TIM-barrel, Rossmann, and Flavodoxin folds.

Specific functional convergences involving different folds

Even on the level of specificity of four-component EC numbers, several enzymatic functions are performed by unrelated structures. Figure 1 shows a dramatic example, two different carbonic anhydrases with the same EC number 4.2.1.1, but with clearly different structures (Kisker *et al.*, 1996). Table 3 shows further examples in a more systematic fashion. Most of these occur in different evolutionary lineages. For instance, the all-alpha Vanadium chloroperoxidase occurs only in fungi, while the alpha/beta non-heme chloroperoxidase occurs only in prokaryotes. Another example is beta-glucanase. It has as many as three different structural representations, from three different fold classes. While it has an all-beta structure in *Bacillus subtilis*, it has an all-

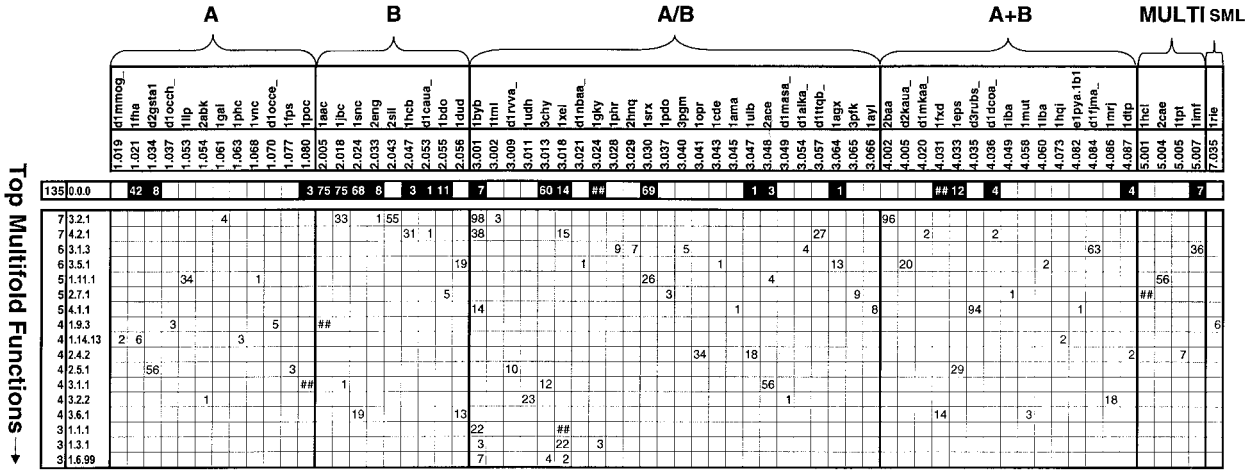


Figure 6. The most versatile functions. Values in the Table denote the number of matches between a particular enzyme category (designated by the first three components of their EC numbers) and a SCOP 1.35 fold (designated by their fold numbers). This figure follows the same conventions described in the legend to Figure 4. The rows are arranged in decreasing order according to the number of different folds with which they are associated (numbers shown in the first column). A hash (#) in any cell indicates that its value is greater than ten.

alpha variant in *Bacillus circulans*, and an alpha/beta structure in tobacco.

Specific functional divergences on same fold

Quite a number of SCOP domains each have sequence similarity with Swissprot proteins of different function. We separated these into cases in which the structural domain has similarity to proteins with different enzymatic functions only and those in which a domain shows homology to both enzymes and non-enzymes (Table 4A and B, respectively). Table 4A includes the well-known lactalbumin-lysozyme C similarity and the well-documented case of homology between an eye-lens structural protein and an enzyme (crystallin and gluthathione S-transferase; Cooper *et al.*, 1993; Qasba & Kumar, 1997). It includes several other notable divergences, such as the one between lysophospholipidase and galectin, and the one between an elastase and an antimicrobial protein (Morgan *et al.*, 1991). Remarkably, of the seven domains in this Table, three belong to the all-beta class.

"Multifunctionality" *versus* *e*-value

Figure 7 shows how the number of "multifunctional" domains, i.e. domains with sequence similarity to proteins with different functions, varies as the function of the stringency of the match score threshold. We used a minimal version of SCOP in which the structures in PDB were clustered into 990 representative domains (see the caption to Figure 6). The Figure shows how the percentage of domains that have sequence similarity to proteins

Table 3. Specific convergences

EC #	Enzymatic function	Fold #1	Dom #1	Swissprot 1	Fold #2	Dom #2	Swissprot 2
1.11.1.10	Chloroperoxidase	3.048.001	d1broa_	PRXC_PSEPY	1.068.001	d1vnc__	PRXC_CURIN
1.15.1.1	Superoxide dismutase	2.001.007	d1srda_	SOD1_ORYSA	4.023.001	d1mnga2	SODM_BACCA
3.1.3.48	Protein-tyrosine phosphatase	3.028.001	d1phr__	PTPA_STRCO	3.029.001	d2hnp__	PYP3_SCHPO
3.1.26.4	Ribonuclease h	3.038.003	d2rn2__	RNH_ECOLI	3.039.001	d1tfr__	RNH_BPT4
3.2.1.4	Endoglucanase	1.061.001	d1cem__	GUN_BACSP	3.001.001	d1ecea_	GUN_BACPO
3.2.1.8	Xylanase	2.018.001	d1yna__	XYN_TRIHA	3.001.001	d2exo__	XYNB_THENE
3.2.1.14	Endochitinase	3.001.001	d1hvq__	CHIA_TOBAC	4.002.001	d2baa__	CHIX_PEA
3.2.1.73	Beta-glucanase*	3.001.001	d1ghr__	GUB_NICPL	2.018.001	d1gbg__	GUB_BACSU
3.2.1.91	Exoglucanase	2.018.001	d1cela_	GUX1_TRIVI	3.002.001	d1cb2a_	GUX3_AGABI
3.5.2.6	Beta-lactamase	5.003.001	d1btl__	BLP4_PSEAE	4.083.001	d1bmc__	BLAB_BACCE
4.2.1.1	Carbonic anhydrase	2.053.001	d1thja_	CAH_METTE	2.047.001	d2cba__	CAHZ_BRARE
5.2.1.8	*Cis-trans* isomerase	4.018.001	d1fkd__	MIP_TRYCR	2.041.001	d2cpl__	CYPR_DROME
5.4.99.5	Chorismate mutase	1.079.001	d1csma_	CHMU_YEAST	4.037.001	d2chsa_	CHMU_BACSU

Explicit enzymatic functions associated with different folds. Of the 13 different enzyme functions listed, eight are hydrolases, five of which belong to the 3.2.1 EC category. One of them, beta-glucanase, is associated with three different folds. Note that most of the enzymes in the Table are associated with folds from different classes. Even when the folds are from the same class, as in the case of protein-tyrosine phosphatases, they are clearly different. Fold numbers are from SCOP 1.35. Domain identifiers are according to the scop syntax: d1pdbc*N*, where "1pdb" is a PDB code, *c* is a chain identifier, and *N* describes if this is the first, second, or only domain in the chain. Thus, d1ggta1 is the first domain in the A chain of 1GGT.

Table 4. Specific divergences

A. *Two different enzymatic functions*

SCOP domain	Fold number	Swissprot 1	EC num 1	Function 1	Swissprot 2	EC num 2	Function 2
d2abk__	1.001.054.001.001.001	END3_ECOLI	4.2.99.18	Endonuclease III	GTMR_METTF	3.2.2.-	Possible G-T mismatches repair enzyme
d1bdo__	1.002.055.001.001.001	BCCP_ECOLI	6.4.1.2	Biotin carboxyl carrier protein of acetyl-Coa carboxylase	BCCP_PROFR	2.1.3.1	Biotin carboxyl carrier protein of methylmalonyl-CoA carboxyl-transferase
d1dhpa_	1.003.001.003.001.004	NPL_ECOLI	4.1.3.3	*N*-Acetylneuraminate lyase subunit	DAPA_BACSU	4.2.1.52	Dihydrodipicolinate synthase
d1hdca_	1.003.018.001.002.005	ENTA_ECOLI	1.3.1.28	2,3 Dihydro-2,3 dihydroxy-benzoate dehydrogenase	ADHI_DROMO	1.1.1.1	Alcohol dehydrogenase 1
d1nipa_	1.003.024.001.005.003	BCHL_RHOCA	1.3.1.33	Protochlorophillide reductase 33 kD subunit	NIFH_THIFE	1.18.6.1	Nitrogenase iron protein
d1gara_	1.003.043.001.001.001	PUR3_YEAST	2.1.2.2	Phosphoribosylglycinamide formyltransferase	PURU_CORSP	3.5.1.10	Formyltetrahydrofolate deformylase
d2dkb__	1.003.045.001.003.001	OAT_RAT	2.6.1.13	Ornithine aminotransferase precursor	GSAB_BACSU	5.4.3.8	Glutamate-1-semialdehyde 2,1-aminomutase 2
d1ede__	1.003.048.001.003.001	DMPD_PSEPU	3.1.1.-	2-Hydroxymuconic semialdehyde hydrolase	HALO_XANAU	3.8.1.5	Haloalkane dehalogenase
d1fua__	1.003.053.001.001.001	ARAD_ECOLI	5.1.3.4	L-Ribulose-5-phosphate 4-epimerase	FUCA_ECOLI	4.1.2.17	L-Fuculose phosphate aldolase
d1lmn__	1.004.002.001.002.010	LCA_RAT	2.4.1.22	Alpha-lactalbumin precursor	LYC1_PIG	3.2.1.17	Lysozyme C-1
d1frva_	1.005.015.001.001.001	FRHG_METVO	1.12.99.1	Coenzyme F420 hydrogenase gamma subunit	MBHS_AZOCH	1.18.99.1	Uptake hydrogenase small subunit precursor

B. *Enzyme and non-enzyme*

SCOP domain	Fold number	Swissprot 1	Enzymatic function	EC number	Swissprot 2	Nonezymatic function
d1gsq_1	1.001.034.001.001.007	GTS2_MANSE	Glutathione *S*-transferase 2	2.5.1.18	SC11_OMMSL	*S*-Crystallin SL11 (major lens polypeptide)
d1lcl__	1.002.018.001.003.003	LPPL_HUMAN	Eosinophil lysophospholipase	3.1.1.5	LEG7_RAT	Galectin-7
d1brbe_	1.002.029.001.002.003	CFAD_RAT	Endogenous vascular elastase	3.4.21.46	CAP7_HUMAN	Azurocidin (antimicrobial, heparin-binding protein)
d1mup__ ..	1.002.039.001.001.007	PGHD_HUMAN	Prostaglandin-D synthase	5.3.99.2	LACC_CANFA	Beta-lactoglobulin III
..d1mup__	1.002.039.001.001.007			QSP_CHICK	Quiescence-specific protein	
d2hhma_ ..	1.005.007.001.002.001	MYOP_XENLA	Inositol mono-phosphatase	3.1.3.25	SUHB_ECOLI	Extragenic suppressor protein SUHB
..d2hhma_	1.005.007.001.002.001	STRO_STRGR	DTDP-glucose synthase	2.7.7.24		
d1isua_	1.007.029.001.001.001	IRO_THIFE	Iron oxidase precursor (FE(II) oxidase)	1.16.3.-	HPIT_RHOTE	High potential iron-sulfur protein (HIPIP)

List of SCOP domains that are each homologous to several Swissprot proteins with significantly different function. In Part A, the domains homologous to proteins with different (in the last three component of EC numbers) enzymatic functions are listed. In most cases, the enzymatic functions remain analogous, as reflected in the names of the enzymes. Part B lists the domains homologous to proteins with both enzymatic and non-enzymatic functions. (See Table 3 for the SCOP domain syntax.)

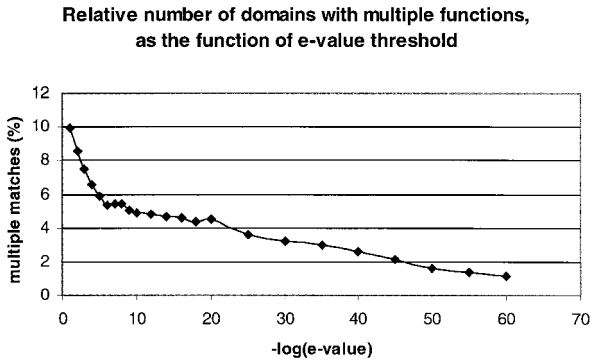


Figure 7. Multi-functionality *versus* *e*-value threshold. The graph shows how the percentage number of multi-functional enzymatic domains varies as the function of the *e*-value threshold. A multi-functional domain occurs when a particular domain in SCOP matches domains in Swissprot with different enzymatic function. For these calculations, we had to use a more minimal version of SCOP than the pdb95d dataset referred to in the methods to prevent double matches, i.e. two SCOP domains matching a single Swissprot domain. The construction of this minimal SCOP was described previously (Gerstein, 1998a,b). Basically, all the domains in SCOP were clustered *via* a multi-linkage approach into 990 representative domains, such that no two domains matched each other with a FastA *e*-value better than .01.

with different functions (in terms of three-component EC numbers) varies with sequence similarity. This decreases approximately monotonically as a function of the exponent of the *e*-value threshold. Interestingly, there is a breaking point around log (*e*-value) $= -5$, as the sharply decreasing number of functions slows down and the matches reach the level of biological significance.

Our graph can be loosely compared with the classic graph by Chothia & Lesk (1986) showing the relation of similarity in structure to that in sequence. It roughly shows the chance of functional similarity (or more precisely the chance of functional difference) with a given level of sequence similarity between an enzyme and a protein of unknown function. For example, with an *e*-value of 10^{-10}, there is only an $\sim$5 % chance that an unknown protein homologous to a certain enzyme has in fact a different function. Moreover, our graph is in excellent agreement with the findings by Russell *et al*. (1998) who also found that the proportion of homologues with different functions is around 10 %. This shows that there is a low chance that a single-domain protein, highly homologous to a known enzyme, has a different function.

Discussion and Conclusions

Overview

We have investigated the relationship between the structure and function of proteins by comparing functionally characterized enzymes in Swissprot with structurally characterized domains in SCOP. It is a timely subject, as the number of three-dimensional protein structures is increasing rapidly and the recent completion of several microbial genomes highlights the need for functional characterization of the gene products and identification of enzymes participating in metabolic pathways (Koonin *et al*., 1998).

We tried to be as objective and as unbiased as possible, taking only enzymes with a single assigned function and only single-domain matches. We ignored Swissprot proteins with dubious or unknown function, or with incomplete sequence. Given these criteria, several tendencies are clear. The alpha/beta folds tend to be enzymes. The all-alpha folds tend to be non-enzymes and the all-beta and alpha + beta folds tend to have a more even distribution between enzymes and non-enzymes.

Our analysis of proteins from yeast and *E. coli* has shown that the functional distribution of folds does not differ greatly from the whole of Swissprot. *E. coli,* however, appears to have somewhat more alpha/beta enzymes and less non-enzymes.

Functional assignment complexities

We identified four specific complexities in our functional assignment worth mentioning.

Firstly, there is not always a one-to-one relationship between gene protein and reaction (Riley, 1998). An enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function. It might be that some of the fold-functions combinations in Figure 2 occur together in multi-domain proteins (which otherwise were not the subject of this survey). An exhaustive screening revealed that only four pairs of folds in Figure 2 were present concurrently in multi-domain proteins. Each of these reduced by one the number of independent fold-function combinations. (The four pairs were as follows, with one representative Swissprot protein in each category, EC numbers in parentheses, and then SCOP fold numbers: PTAA_ECOLI (2.7.1) has 4.049 and 2.055 folds, TRP_COPCI (4.2.1) has 3.057 and 4.005 folds, URE1_HELFE ([3.5.1) has 4.005 and 2.056 folds, while XYNA_RUMFL (3.2.1) has 2.018 and 3.001 folds.)

Secondly, The functions associated with similar structures often turn out to be analogous, even if they show significant difference in their EC numbers. For example, acetyl-CoA carboxylase and methylmalonyl-CoA carboxyltransferase enzymes are both actually part of enzyme complexes in which they perform the same function, acting as enzyme carriers. This similarity is not reflected in their EC classification numbers (6.4.1.2 and 2.1.3.1, respectively).

Thirdly, there are clearly some drawbacks to the EC system. The EC system is a classification of

reactions, not underlying biochemical mechanisms. An enzyme classification system based explicitly on reaction mechanism (e.g. ''involves pyridoxal phosphate'' or ''involves Ser as a nucleophile'') might also prove interesting to compare with protein structure. Alternatively, one based on pathways might be worthwhile since, as pointed out by Martin *et al*. (1998), ''it may be that more significant relationships occur within pathways, where the substrate is successively transferred from enzyme to enzyme along the pathway, requiring similar binding sites at each stage''.

Finally, in all of Swissprot the majority of the 101 folds with only non-enzymatic functions probably have several functions, but we were not able to consider them separately here, lacking a general protein function classification system for non-enzymes. Such a system is not easy to derive. For instance, if we took only the first three words of all the description lines in Swissprot, we would end up with about 10,000 different protein functions (besides enzymes). An approximate solution to this problem is offered by a recent work that has classified 81 % of Swissprot into one of three broad categories in an automated fashion (Tamames *et al*., 1997). However, one way we did tackle this problem was by focussing on the yeast genome for which there are a number of overall functional classification systems. This work showed that the preferred association of folds with certain functions occurs for non-enzymes as well as enzymes. Furthermore, the results for the highly conserved COGs would be expected to be exactly the same in other genomes.

Biases

Our results are undoubtedly affected to some degree by the biases inherent in the databanks, e.g. towards mammalian, medically relevant proteins and towards proteins that easily crystallize. Such biases probably result in the higher representation of enzymes in the structural databases, in the PDB and therefore in SCOP. This might be the cause of the higher occurrence of alpha/beta proteins in our tables and the higher density of matches in this class.

One interesting question related to biases is whether looking only at individual genomes instead of the whole database will give different results. Our results for yeast suggest that it is not necessarily the case.

Comparison with Martin *et al*. (1998)

Martin *et al*. (1998) performed a similar analysis to the one described here. One of the conclusions of their careful study was that there was no relationship between the top-level CATH classification and the top-level EC class. This seems to be at odds with our results. However, we have found the conclusions to be consistent. There are a number of reasons for this.

Firstly, Martin *et al*. (1998) tabulate statistics on only the proteins in the PDB. They found a clear alpha/beta preference for proteins in the oxido-reductase, transferase, and hydrolase categories (EC 1-3), but for the lyase, isomerase, and ligase categories (EC 4-6) they observe different tendencies. However, they did not have sufficient counts to establish statistical significance for this latter finding. (This is basically what we observe in Figure 4(b)). Because in our analysis we use all of Swissprot and we tabulate our statistics a little differently (in terms of combinations), we get more ''counts'' than Martin *et al*. (1998). Thus, we are able to argue that the different distribution of fold function combinations observed for lyases, isomerases, and ligases are significant. This is borne out by the chi-squared statistics at the end of Table 2.

Secondly, Martin *et al*.'s ''no-relationship'' conclusion applies only to comparisons between the different enzyme classes. However, we find our largest differences when comparing non-enzymes to enzymes and also comparing between the various types of non-enzymes.

Finally, the CATH classification that Martin *et al*. use has only three classes in its top-most level. In contrast, SCOP has six top classes (Table 1). While this larger number of categories does tend to degrade our statistics somewhat, it also highlights some differences that cannot be observed in terms of the CATH classes alone, e.g. we find clear differences between alpha + beta and alpha/beta proteins and also between small proteins and all others.

Apparently high occurrence of convergent evolution

Note that the Table in Figure 2 is not square: it has more folds than functions. This shape leads to a number of interesting conclusions. The 331 fold-function combinations we observe for 229 folds and 92 functions imply that there are 1.2 functions per fold and 3.6 folds per function. However, these numbers are somewhat skewed by the large number of folds (101) associated only with the single non-enzymatic function. If we exclude these, we get 128 ''enzyme-related'' folds, which are, in turn, associated with 230 $(=331-101)$ different fold-function combinations. This implies that for the enzyme-related folds there are on average 1.8 functions per fold and 2.5 folds per function (230/128 and 230/92). The larger number of folds per function than functions per fold seems to suggest that nature tends to reinvent an enzymatic function (i.e. convergent evolution) more often than modify an already existing one (i.e. functional divergence).

How can we explain this? Firstly, 1.8 is a lower estimation for the number of functions per fold as the non-enzymatic functions were bundled into one group here. Secondly, there are several examples of functional divergence for a fold within one three-component enzyme category that are not

reflected in our Tables. For instance, the 1.1.1 category has 248 different enzymes, which all share the same fold. Thirdly, the results in this paper were derived from databases comprised of data from several organisms. It is quite possible that within one organism, functional divergence is more prevalent than convergent evolution.

Superfolds and superfunctions

Are functions more diverse for the more common folds? To some degree this brings up a ''chicken-and-the-egg'' issue. Do folds have more functions because they occur more often or is it the other way around? The commonness of a fold is often quantified by the number of non-homologous sequence families accommodated by the fold, and folds accommodating many families of diverse sequences have been dubbed ''superfolds'' (Orengo *et al*., 1993). We find that there seems to be a loose connection between the number of diverse sequence families associated with a particular fold (in SCOP) and the functional diversity of that fold. For instance, the top superfold is the TIM-barrel; it also has the most functions associated with it (15 different enzymatic functions as shown in Figure 4). On the other hand, there are exceptions: the alpha/beta hydrolases and the Rossmann fold are both associated with 22 sequence families in SCOP, but while the former has eight different enzymatic functions, the latter has only three.

Finally, while there is a high incidence of particular functions with many folds (''superfunctions''), as well as folds with many functions, the distribution of superfunctions appears to be more uniform and less concentrated on a few exceptionally versatile individuals than is the case for folds. That is, comparing Figures 3 and 4 one can see that the top nine most versatile functions are associated with five to seven folds while the top nine most versatile folds carry out from six to as many as 16 functions. This last value is for the TIM-barrel and underscores the uniqueness of this fold as a generic scaffold (see Figure 1 for an illustration of this fold).

Why folds are associated with functions: chemistry *versus* history

Why is a certain fold chosen to carry out a particular function? It is, of course not possible to answer this question definitively at present. However, there are two broad themes that emerge from our analysis. The first is favorable chemistry. Perhaps the TIM-barrel design simply provides a ''more efficient'' scaffold for enzyme reactions so that is why it is so prevalent. Another factor is history. Perhaps the association between a particular fold and its function reflects a particular ''accident'' that took place at the beginning of cellular evolution. However, once this choice was made it was impossible to undo even if other folds would be more chemically suitable. This could be the situation for the ribosomal proteins (and is borne out by the results of Figure 4(d)).

Materials and Methods

Sequence matching to swissprot

All the protein sequences in Swissprot 35 were compared with all the protein domain sequences in SCOP 1.35 by standard database search programs (WU-BLAST; Altschul *et al*., 1990). The following five criteria were used in the searches: (1) At least three of the four components of the EC number are assigned in the DE line of the Swissprot entries. (2) Fragments in Swissprot were excluded (this affected about 10 % of the entries). (3) For WU-BLAST searches an *e*-value threshold of .0001 was used, unless stated otherwise. (4) Only ''monoenzymes'', i.e. proteins with only one enzymatic function, were considered. This excluded less than 0.5 % of the Swissprot enzymes. (5) Only single-domain matches with Swissprot proteins were taken into consideration. This means those proteins that had a match with a SCOP domain covering most of the Swissprot protein. Specifically, we required that less than 100 amino acid residues be left uncovered in the Swissprot entry by a match. We are aware that this is only an approximation, as there are domains with less than 100 amino acid residues; however it is considerably less than the average length of a SCOP domain (163 residues) and seems to be a reasonable threshold in an automated approach.

All the searches were repeated using FASTA with an *e*-value threshold of .01 (Pearson, 1998; Pearson & Lipman, 1988). The results obtained by the two different comparison programs were in agreement with each other. That is, the FASTA searches did not result in any new combinations of folds and enzymatic functions (a new dot in Figure 1), and therefore are not shown.

Sequence matching to the yeast genome

To get as great a coverage of the yeast genome as possible, we did a sequence comparison for *just* Figure 4 using an altered protocol. We first ran the PDB against the yeast genome using FASTA and kept all matches with a better than 0.01 *e*-value (Pearson, 1998; Pearson & Lipman, 1988). Then, to increase our number of matches further we used the PSI-blast program (Altschul *et al*., 1997). This program is somewhat more complex to run than FASTA, involving embedding the yeast genome in NRDB and running PDB query sequences against it in an iterative fashion, adding the matches found at each round to a growing profile. We used the PSI-blast parameters adapted from Teichmann *et al*. (1998): an *e*-value threshold of .0005 to include matches in the profile and iteration of up to 30 times or to convergence. We did not continuously parse the output and accepted matches at the final iteration that had *E*value scores better than .0001. The number of iteration to convergence varies depending on the PDB domains being run. Runs that take many iterations such as those for the immunoglobulin superfamily take quite a long time (up to 30 minutes on DEC 500 MHz workstation) and create large output files. In total, PSI-blast finds many more matches than either FASTA or WU-BLAST. However, it has problems with certain small and compositionally biased proteins. We used FASTA for these and also tried to remove compositional bias

through running the SEG program with standard parameters (Wootton & Federhen, 1996).

How the structural classifications were used: SCOP and CATH

SCOP hierarchically clusters all the domains in the PDB database, assigning a five-component number to each domain (Murzin *et al.*, 1995). The first component in the SCOP numbers denotes the structural class to which the domain in question belongs. The second component of the SCOP numbers designates the fold type of the domain. There are altogether 361 different fold types in SCOP 1.35. The six SCOP classes used in this survey are listed in Table 1B.

In this study, a 95% non-redundant subset of SCOP was used, i.e. all pairs of domains had less than 95% sequence homology. This set is denoted pdb95d and is available from the SCOP website (scop.mrc-lmb.cam.ac.uk). We used version 1.35, which had 2314 protein domains. (The yeast analysis used a more recent version of SCOP, 1.38, which had 3206 domains.)

The CATH classification classifies structures in analogous fashion to SCOP (Orengo *et al.*, 1997). However, the exact structure of the classification is not the same, with an additional architecture level inserted between the top-level class and the fold-level. In our use of the classification, we created a limited mapping table that associated each SCOP domain in pdb95d with its corresponding classification in CATH 1.4. This was not always possible to do unambiguously. As a result, we left out the ambiguous matches from the statistics.

How the functional classifications were used: ENZYME, COGS, and MIPS

The EC numbers of enzymes are composed of four components (Barrett, 1997): (1) The first component shows to which of the six main divisions the enzyme belongs; (2) the second figure indicates the subclass (referring to the donor in oxidoreductases or the group transferred in transferases, or the affected bond in hydrolases, lyases or ligases); (3) the third figure indicates the sub-subclass (e.g. indicating the type of acceptor in oxidoreductases), and (4) the fourth figure gives the serial number of the enzyme in its sub-subclass. The six main divisions are listed in Table 1A.

In the analysis of all of Swissprot, when we counted the number of non-enzymatic matches, all the proteins called 'HYPOTHETICAL' and all the proteins having an '-ase' word ending but lacking an EC number in their description were excluded, because of their functional ambiguity. For relating the sequence matches of the yeast genome to the EC system, we used essentially the same criteria as we did for all of Swissprot (see above): single-domain, monoenzyme matches with at least a three-component EC number.

The COGs and especially the MIPS classifications are a bit more complex than the EC system in that they include non-enzymes as well as enzymes (Tatusov *et al.*, 1997; Koonin *et al.*, 1998; Mewes *et al.*, 1997). They often associate multiple functions or roles to a given yeast ORF. This happens for more than a third of the yeast ORFs with MIPS. In this case, if we could clearly show a PDB match was associated with a single functional domain we made only that pairing. Otherwise we associated all the functions assigned to a given PDB match to its respective fold.

Availability of results over the internet

A number of detailed tables relevant to our study will be made available over the Internet at http://bioinfo.mbb.yale.edu/genome/foldfunc, in particular, a ''clickable'' version of Figure 1 and large data files giving all the fold assignment and fold-function combinations for Swissprot and yeast.

Acknowledgments

We thank the Donaghue Foundation and the ONR for financial support (grant N000149710725). We thank Ted Johnson for help with the minimal version of the SCOP database.

References

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997b). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304-308.

Bairoch, A. (1996). The ENZYME data bank in 1995. *Nucl. Acids Res.* **24**, 221-222.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Barrett, A. J. (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur. J. Biochem.* **250**, 1-6.

Bork, P. & Eisenberg, D. (1998). Deriving biological knowledge from genomic sequences. *Curr. Opin. Struct. Biol.* **8**, 331-332.

Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.* **18**, 313-318.

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2**, 31-40.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.

Chen, L., DeVries, A. L. & Cheng, C. H. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl Acad. Sci. USA,* **94**, 3817-3822.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Cooper, D. L., Isola, N. R., Stevenson, K. & Baptist, E. W. (1993). Members of the ALDH gene family are lens and corneal crystallins. *Advan. Exp. Med. Biol.* **328**, 169-179.

Coque, J. J., Liras, P. & Martin, J. F. (1993). Genes for a beta-lactamase, a penicillin-binding protein and a transmembrane protein are clustered with the cephamycin biosynthetic genes in *Nocardia lactamdurans*. *EMBO J.* **12**, 631-639.

Corpet, F., Gouzy, J. & Kahn, D. (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323-326.

des, Jardins M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB,* **5**, 92-99.

Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15-18.

Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H. & Pongor, S. (1997). The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nucl. Acids Res.* **25**, 240-243.

Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626-628.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779-790.

Gerstein, M. (1997). A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Design,* **3**, 497-512.

Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.

Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277-304.

Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA,* **94**, 11911-11916.

Hellinga, H. W. (1997). Rational protein design: combining theory and experiment. *Proc. Natl Acad. Sci. USA,* **94**, 10015-10017.

Hellinga, H. W. (1998). Computational protein engineering. *Nature Struct. Biol.* **5**, 525-527.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309-312.

Hodges, P. E., Payne, W. E. & Garrels, J. I. (1998). The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucl. Acids Res.* **26**, 68-72.

Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316-319.

Ibba, M., Bono, J. L., Rosa, P. A. & Soll, D. (1997a). Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA,* **94**, 14383-14388.

Ibba, M., Morgan, S., Curnow, A. W., Pridmore, D. R., Vothknecht, U. C., Gardner, W., Lin, W., Woese, C. R. & Soll, D. (1997b). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science,* **278**, 1119-1122.

Karp, P. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics,* **14**, 753-754.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998). EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50-53.

Kisker, C., Schindelin, H., Alber, B. E., Ferry, J. G. & Rees, D. C. (1996). A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon *Methanosarcina thermophila*. *EMBO J.* **15**, 2323-2330.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7**, 757-763.

Koonin, E. V. & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.* **244**, 125-132.

Koonin, E. V., Tatusov, R. L. & Galperin, M. Y. (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**, 355-363.

Kraulis, P. J. (1991). MOLSCRIPT-a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. & Thornton, J. M. (1998). Protein folds and functions. *Structure,* **6**, 875-884.

Marvin, J. S., Corcoran, E. E., Hattangadi, N. A., Zhang, J. V., Gere, S. A. & Hellinga, H. W. (1997). The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl Acad. Sci. USA,* **94**, 4366-4371.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. & Zollner, A. (1997). Overview of the yeast genome. *Nature,* **387**, 7-65.

Morgan, J. G., Sukiennicki, T., Pereira, H. A., Spitznagel, J. K., Guerra, M. E. & Larrick, J. W. (1991). Cloning of the cDNA for the serine protease homolog CAP37/azurocidin, a microbicidal and chemotactic protein from human granulocytes. *J. Immunol.* **147**, 3210-3214.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29-34.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identifying and classifying protein fold families. *Protein Eng.* **6**, 485-500.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-259.
Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.
Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA,* **85**, 2444-2448.
Qasba, P. K. & Kumar, S. (1997). Molecular divergence of lysozymes and alpha-lactalbumin. *Crit. Rev. Biochem. Mol. Biol.* **32**, 255-306.
Riley, M. (1997). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl. Acids Res.* **25**, 51-52.
Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211-1227.
Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.
Seery, L. T., Nestor, P. V. & FitzGerald, G. A. (1998). Molecular evolution of the aldo-keto reductase gene superfamily. *J. Mol. Evol.* **46**, 139-146.
Selkov, E., Galimova, M., Goryanin, I., Gretchkin, Y., Ivanova, N., Komarov, Y., Maltsev, N., Mikhailova, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L. & Selkov, E., Jr. (1997). The metabolic pathway collection: an update. *Nucl. Acids Res.* **25**, 37-38.
Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405-420.
Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.
Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science,* **278**, 631-637.
Teichmann, S., Park, J. & Chothia, C. (1998). Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA,* **95**, 14658-14663.
Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.

Edited by G. Vontleijne

(Received 16 November 1998; received in revised form 1 March 1999; accepted 1 March 1999)